Daniel Fragoso Alvarado

Tema 4 — Actividad no. 1
Cuestionario

7 de septiembre, 2022

1. ¿Qué son los corpus y bajo qué criterios se pueden clasificar?

   Los corpus son una serie de ejemplos lingüísticos recopilados de diferentes fuentes sobre algo que se desea estudiar dentro de la lengua; hasta antes de mediados del siglo XX, los datos en lingüística eran una mezcla de datos observados y ejemplos inventados.

   Y a pesar de la negativa de Chomsky del uso de corpus, se ha demostrado que los corpus son muy útiles en una variedad de áreas de la lingüística, brindando información en áreas tan diversas como la lingüística contrastiva, el análisis del discurso, el aprendizaje de idiomas, semántica, sociolingüística y lingüística teórica. Según el tipo de información lingüística que se estructura se clasifica el tipo de corpus.

2. Visita el sitio de un corpus de tu elección de los listados en la sección 4.5 y, siguiendo la tabla de la sección 4.1.2 de la presentación, caracterízalo.

   Para este ejercicio se seleccionó *Corpus del Español de Mark Davies*, dentro de la plataforma existen a su vez 4 corpus más, seccionamos; vamos a elegir el corpus de *Genre/Historical*. Este corpus contiene más de 100 millones de palabras en más de 20.000 textos en español desde los años 1200 al 1900. Los 20 millones de palabras de texto de la década de 1900 se equilibran entre los géneros hablado, ficción, periódico y académico. Este corpus le permite comparar entre géneros y períodos de tiempo.

   Entonces, siguiendo la clasificación que se sigue en la presentación, es un corpus de tipo textual, literario: Narrativa e informativo: periodísticos, académicos. Con textos sincrónicos de tipo histórico, y diacrónico de forma de periodísticos.

3. ¿Cuáles son los tres tipos de información que contiene un corpus además de los textos o transcripciones? Descríbelos brevemente.

   Metadatos, Marcado textual y Anotación lingüística:

   - Metadatos: Es información sobre el texto *per se*. Es decir nos pueden decir información sobre el autor, año de publicación, idioma, editorial, etc. Los metadatos se pueden codificar en el texto del corpus, o en un documento o base de datos separados.
   - Marcado Textual: El marcado textual codifica información dentro del texto distinta de las palabras reales. Por ejemplo, en un texto impreso, el marcado textual normalmente se usaría para representar el formato del texto, como dónde comienzan y terminan las cursivas. En los corpus orales transcritos, la información transmitida por los metadatos y el marcado textual puede ser muy importante para el análisis de la transcripción.
   - Anotación lingüística: Es la codificación de información lingüística dentro de un texto del corpus de tal manera que podamos recuperar ese análisis de manera sistemática y precisa más adelante; cuando se hace esto.

     La anotación suele utilizar las mismas convenciones de codificación que el marcado textual.

4. ¿Por qué resulta tan importante la consistencia en la anotación de un corpus?

   De no existir la anotación del corpus, el potencial de inconsistencias en el análisis de un texto se incrementa de forma considerable, esto no quiere decir que las anotaciones sean 100 % confiables, y que no vaya a existir estas inconsistencias si el corpus usado tienen anotaciones, sin embargo sí significa que se tiene la posibilidad de encontrar el error y solucionarlo; lo que seria prácticamente imposible si el corpus o tiene anotaciones.

5. ¿Cuáles son las principales características de las cuatro generaciones de concordancers?

   - **1ra Generación**: Los concordancers de primera generación generalmente se guardaban en una computadora central y se usaban en un solo sitio. Los equipos de investigación individuales construirían su propio sistema de concordancia y lo usarían en los datos a los que tenían acceso localmente.
   - **2da Generación**: Los concordancers de segunda generación fueron habilitados por la difusión de máquinas (compatibles con IBM). Se hizo posible escribir y distribuir un concordancer para esa plataforma con una alta probabilidad de que el programa se pudiera usar de inmediato en la máquina del destinatario.
   - **3ra Generación**: La tercera generación de software de concordancia incluye sistemas. Estos concordancers pudieron manejar grandes conjuntos de datos en una computadora. Además, habían incluido en ellos una gama más amplia de herramientas que las que habían estado disponibles anteriormente en los paquetes de concordancia, y dieron acceso a algunos análisis estadísticos significativos que iban más allá de la estadística descriptiva. Finalmente, apoyaron efectivamente una variedad de sistemas de escritura.
   - **4ta Generación**: La cuarta generación de concordancers es sorprendentemente similar a la tercera. De hecho, han surgido concordancers de cuarta generación, no para ampliar el rango de análisis disponibles, sino para abordar tres problemas completamente diferentes: el poder limitado de las PC de escritorio, los problemas que surgen de los sistemas operativos de PC no compatibles y las restricciones legales en la distribución. de corpus

     Sin embargo, las herramientas de análisis de corpus de cuarta generación son incluso más fáciles de usar y potentes que las herramientas de tercera generación.

6. ¿Cuáles son las tres herramientas más usadas en el análisis de corpus y qué utilidad tienen?

   - Recopilación: Como su nombre lo dice, ayudan a la recopilación de texto, de diferentes fuentes según sea el caso.
   - Etiquetado: Sirve para generar el etiquetado del contenido en el texto.
   - Análisis: Sirven para generar análisis sobre l¡el texto, por ejemplo estadístico .

7. McEnery & Hardy mencionan una idea alternativa a usar software open-source para el análisis de corpus. ¿Cuál es esta y cuáles son sus ventajas?

   Proponen que cualquier persona que trabaje con corpus podría crear un programa de concordancia existente cuyo código fuente está disponible abiertamente, posteriormente el programa se lanza a todos los usuarios del sistema. Esto reduce la cantidad de trabajo que se debe realizar para ponerlo en funcionamiento y significa que se evita el trabajo extra en cada trabajo. Si este enfoque se adoptara en el contexto de los oncordancers de cuarta generación, entonces bien podría surgir una solución integral a los problemas que enfrentan los desarrolladores y usuarios de herramientas de corpus.

8. ¿Cuáles son las desventajas que tiene esta alternativa?

   El principal problema es que la mayoría de los programas de búsqueda y recuperación de corpus son sistemas enormemente complicados, tan complicados que puede ser muy difícil para cualquiera que no sea su mantenedor principal comprender su funcionamiento interno lo suficientemente bien como para comenzar a programar extensiones.

9. ¿Qué es la frecuencia relativa de una palabra en un corpus y cómo se calcula?

   A veces, como es el caso aquí, es posible que el porcentaje no transmita de manera significativa la frecuencia de uso de la palabra, por lo que podríamos producir una frecuencia relativa/normalizada, que responde a la pregunta "¿con qué frecuencia podemos suponer que veremos la palabra por cada x palabras del texto en ejecución?":

$$nf = (\text{número de ejemplos de la palabra en todo el corpus} \div \text{tamaño del corpus}) \times (\text{base de normalización})$$

10. ¿Cuáles son los principales tests de significatividad y por qué resultan relevantes?

    Los dos usos más comunes de las pruebas de significación son el cálculo de palabras clave (o etiquetas clave) y el cálculo de colocaciones. Para extraer palabras clave, necesitamos probar la significación de cada palabra que aparece en un corpus, comparando su frecuencia con la de la misma palabra en un corpus de referencia. Cuando buscamos las colocaciones de una palabra, probamos la importancia de la frecuencia de co-ocurrencia de esa palabra y todo lo que aparece cerca de ella una o más veces en el corpus. Ambos procedimientos suelen implicar, entonces, la realización de muchos miles de pruebas de significancia.